Ein System zur Interpretation einer Äußerung: Realisierung

S. Kunzmann*, T. Kuhn, E. Nöth, G. Stallwitz

Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen-Nürnberg
*IBM Deutschland GmbH, Wissenschaftliches Zentrum Heidelberg

1 Einführung

Das Spracherkennungs- und Dialogsystem EVAR (Erkennen, Verstehen, Antworten, Rückfragen) [Nie91] ist ein Auskunftssystem für den Diskursbereich *"Intercity-Zugauskunft"*. Der Benutzer soll mit dem System in einem telefonischen Auskunftsdialog Informationen über Intercity-Verbindungen und Fahrpreise einholen können. Dies erfordert sowohl das Erkennen als auch das Interpretieren von Benutzeräußerungen. Im folgenden wird die Struktur und das Zusammenwirken einer Arbeitsumgebung beschrieben, welche es erlaubt, die erste Version des Spracherkennungs- und Dialogsystems EVAR zu realisieren.

2 Anforderungen an eine interaktive Arbeitsumgebung

Die Gesamtanalyse einer Äußerung innerhalb von EVAR umfaßt eine Vielzahl von sequentiell ablaufenden Einzelschritten bestehend aus den Analyseebenen:

- Aufnahme und Digitalisierung eines Sprachsignals
- Merkmalberechnung
- Lautkomponentenklassifikation
- Lautsegmentierung und -klassifikation
- Worthypothesengenerierung
- Linguistische Analyse inklusive Antwortgenerierung

Um einem Anwender eine einfache Durchführung der Gesamtanalyse einer Äußerung zu ermöglichen, wurde bei der Realisierung der Arbeitsumgebung eine einfache Bedienoberfläche von der Aufnahme bis zur Antwortgenerierung geschaffen. Darüber hinaus sollten sämtliche zur Verfügung stehenden Methoden in das Gesamtsystem integriert werden. Dies ermöglicht es einerseits Wechselwirkungen zwischen den verschiedenen Methoden untereinander zu untersuchen und andererseits mit alternativen Erkennungs- oder Analysemethoden im Rahmen des Gesamtsystems zu experimentieren. Für weitere Untersuchungen werden dem Anwender der Arbeitsumgebung Hilfsmittel zur Verfügung gestellt um die generierten Ergebnisse interaktiv zu evaluieren bzw. automatisch auszuwerten.

Die Bedienoberfläche der Arbeitsumgebung wurde als Baum organisiert. Ausgehend vom Hauptmenü gelangt man wahlweise in eines der folgenden Menüs:

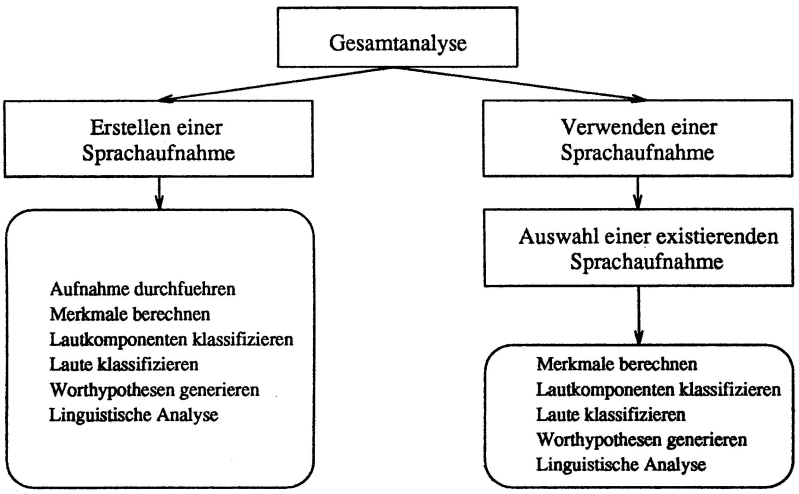

Abbildung 1: Struktur des Menüs *Gesamtanalyse*

- *Gesamtanalyse*
- *Schrittweise Analyse*
- *Parametereinstellung*

Im Menü *Gesamtanalyse* laufen die verschiedenen Verfahren in einer definierten Reihenfolge sequentiell ab, wobei gewählt werden kann, ob man auf eine "Konserve", d.h auf eine bereits existierende Sprachaufnahme zurückgreifen will, oder ob die Gesamtanalyse mit der Aufnahme einer neuen Äußerung beginnen soll. Die Struktur der Gesamtanalyse ist in Abbildung 1 dargestellt, wobei die verwendeten Methoden in [Kuh91a] beschrieben sind. Im Menü *Schrittweise Analyse* hat man die Möglichkeit, die in der Gesamtanalyse verwendeten Methoden unabhängig voneinander aufzurufen oder auch alternative Verfahren einzusetzen (z.B. Methoden zur Verifikation von Worthypothesen). Damit können Experimente mit unterschiedlichen Algorithmen durchgeführt und die Ergebnisse im Rahmen des Gesamtsystems ausgewertet werden. Durch das Menü *Parametereinstellung* lassen sich die Parameter der verschiedenen, in die Arbeitsumgebung integrierten, Verfahren verändern. Unter anderem besteht die Möglichkeit verschiedene Merkmalssätze auszuwählen, den zu Grunde gelegten Wortschatz bei der Erkennung neu zu definieren oder direkt die Parameter der einzelnen Algorithmen zu verändern. Beispielsweise kann man auf diese Weise für die Worthypothesengenerierung die Anzahl der zu generierenden Worthypothesen variieren oder die stichprobenabhängigen Parameter (HMM-Parameter, Dimensionierung des Normalverteilungsklassifikators) in Abhängigkeit der beim Training verwendeten Stichprobe wählen [Kuh91b]. Falls beim Parametertraining genau ein Sprecher verwendet wird, kann auf diese Weise eine sprecherabhängige Version des Gesamtsystems realisiert werden.

3 Hardware-Konfiguration von EVAR

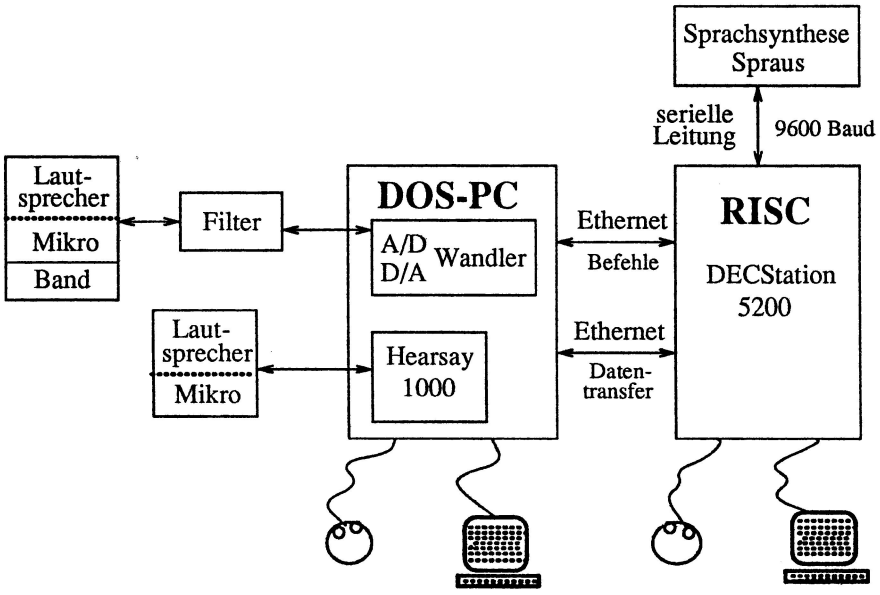

Abbildung 2: Hardware-Konfiguration von EVAR

Eine der Hauptzielsetzungen für EVAR und damit auch für die Arbeitsumgebung ist die Unabhängigkeit von Spezialhardware. Dementsprechend ist die Hardware-Konfiguration ausgelegt (Abbildung 2). Zur Verfügung steht eine RISC-Workstation (DECStation 5200, 25 mips) mit Ultrix, ausgestattet mit 32 MB Hauptspeicher. An dieser Workstation ist über eine serielle Leitung ein Sprachsynthesegerät (deutsche Ausgabe) von Daimler-Benz angeschlossen. Die Sprachaufnahme erfolgt an einem PC-AT unter MSDOS, der über Ethernet mit der Workstation verbunden ist. In dem PC ist eine A/D-D/A-Karte (DT2801-A) sowie eine Ethernet Karte integriert. Für die Durchführung der Aufnahme steht ein Filter, ein Mikrofon und ein analoges Kassettengerät zur Verfügung. Darüber hinaus ist an den PC ein weiteres Synthesegerät (Hearsay 1000, englische Ausgabe) für Demonstrationszwecke angeschlossen.

Nachdem zum Führen eines Dialogs die Sprachdaten in Echtzeit von der A/D-D/A-Karte an die Erkennungsalgorithmen übertragen werden müssen, wurde eine flexible Kommunikation zwischen dem PC und der Workstation basierend auf dem *Client-Server-Modell* unter Verwendung des TCP/IP-Protokolls realisiert. Die Workstation (*Client*) sendet Befehle über die TCP/IP-Verbindung an den PC (*Server*) um Aktionen zu starten, die die Programmierung des PC's bzw. der A/D-D/A-Karte betreffen. Beispiele für mögliche Aktionen sind Befehle zum Aufnehmen, Wiedergeben oder Darstellen einer Äußerung. Der PC übernimmt als *Server* den Datenaustausch zwischen der A/D-D/A-Karte und der Workstation über TCP/IP. Tests haben ergeben, daß bei normaler Auslastung der Workstation bis zu 64

KB pro Sekunde übertragen werden können. Bei einer Abtastrate von 16 kHz fallen bei einer Auflösung von bis zu 16 Bit 32 KB Daten an. Dies entspricht einer Übertragungsgeschwindigkeit von 1/2-facher Echtzeit, so daß noch genügend Spielraum für eine höhere Auslastung der Workstation zur Verfügung steht.

4 Beschreibung eines Interpreters für eine interaktive Arbeitsumgebung

Für den Aufbau einer interaktiven Arbeitsumgebung wurde eine einfache Programmiersprache zur Erzeugung von Kommandofenstern oder Menüs - in Anlehnung an die UNIX-Terminologie im folgenden *Scripten* genannt - im X-Window System realisiert. Basierend auf dieser Programmiersprache schreibt ein Benutzer mit einem Standardtexteditor ein *Script*, in welchem Darstellungs- (graphische Attribute, Fonts, usw.) und Aktionsanweisungen stehen. Durch die Struktur der *Scripts* kann eine Bedienerführung vorgenommen werden, wobei die Interpretation der internen Struktur auf Grund von Benutzereingaben erfolgt. Die Menüs erlauben die Definition einer Benutzerführung, wohingegen den Buttons in Kommandofenstern Aktionen zugeordnet sind (z.B. Merkmalberechnung, Setzen bestimmter Parameter). An jeden Button können darüber hinaus online-Hilfemeldungen gebunden werden, die den Anwender bei der Benutzung der Arbeitsumgebung unterstützen sollen. Ein weiteres, wesentliches Merkmal des realisierten Interpreters erlaubt die Verwendung sogenannter *symbolischer Links*. Damit müssen mehrfach benötigte Aktionen nur einmal formuliert werden. Der Aufbau der 3000 Programmzeilen umfassenden Arbeitsumgebung für EVAR benötigt auf einer DECStation 5200 ca. 2 Sekunden.

5 Zusammenfassung

Vorgestellt wurde die realisierte Arbeitsumgebung des Spracherkennungs- und Dialogsystems EVAR. Die Aufnahme einer Äußerung erfolgt an einem PC-AT während die weitere Analyse von der Merkmalberechnung bis zur Antwortgenerierung an einer RISC-Workstation unter dem Betriebssystem Ultrix stattfindet. Um den Datenaustausch zwischen PC und Workstation zu ermöglichen wurde ein *Client-Server-Modell* unter Verwendung des TCP/IP-Protokolls realisiert. Eine benutzerfreundliche, auf dem X-Window-System basierende Bedienoberfläche erlaubt eine einfache Systemkonfiguration sowie ein interaktives Untersuchen der erzeugten Daten.

Literatur

[Kuh91a] T. Kuhn, S. Kunzman, F. Kummert, M.Mast, H. Niemann, E. Nöth, R. Prechtel, S. Rieck, A. Reißser, G. Sagerer, E. Schukat-Talamazzini, J. Unglaub: *Ein System zur Interpretation einer Äußerung: Methoden.* In *Fortschritte der Akustik-DAGA '91*, 1991.

[Kuh91b] T. Kuhn, S. Kunzmann, E. Noeth, S. Rieck, E. Schukat-Talamazzini: *Iterative optimization of the Data Driven Analysis in Continuous Speech.* In *NATO ASI Speech Recognition and Understanding: Recent Advances, Trends and Applications*, Springer-Verlag, Berlin, erscheint 1991.

[Nie91] H. Niemann: *The combination of word recognition and linguistic processing in speech understanding.* In *NATO ASI Speech Recognition and Understanding: Recent Advances, Trends and Applications*, Springer-Verlag, Berlin, erscheint 1991.